RemoteCLIP: A Vision Language Foundation Model for Remote Sensing

Year: 2023 | Citations: 409 | Authors: F. Liu, Delong Chen, Zhan-Rong Guan, Xiaocong Zhou, Jiale Zhu

---

Abstract

General-purpose foundation models have led to recent breakthroughs in artificial intelligence (AI). In remote sensing, self-supervised learning (SSL) and masked image modeling (MIM) have been adopted to build foundation models. However, these models primarily learn low-level features and require annotated data for fine-tuning. Moreover, they are inapplicable for retrieval and zero-shot applications due to the lack of language understanding. To address these limitations, we propose RemoteCLIP, the first vision-language foundation model for remote sensing that aims to learn robust visual features with rich semantics and aligned text embeddings for seamless downstream application. To address the scarcity of pretraining data, we leverage data scaling which converts heterogeneous annotations into a unified image-caption data format based on box-to-caption (B2C) and mask-to-box (M2B) conversion. By further incorporating unmanned aerial vehicle (UAV) imagery, we produce a $12\times $ larger pretraining dataset than the combination of all available datasets. RemoteCLIP can be applied to a variety of downstream tasks, including zero-shot image classification, linear probing, k-NN classification, few-shot classification, image-text retrieval, and object counting in remote sensing images. Evaluation of 16 datasets, including a newly introduced RemoteCount benchmark to test the object counting ability, shows that RemoteCLIP consistently outperforms baseline foundation models across different model scales. Impressively, RemoteCLIP beats the state-of-the-art (SOTA) method by 9.14% mean recall on the RSITMD dataset and 8.92% on the RSICD dataset. For zero-shot classification, our RemoteCLIP outperforms the contrastive language image pretraining (CLIP) baseline by up to 6.39% average accuracy on 12 downstream datasets.